PHILOLOGIA 42 (2011) 89-105

エクセルを用いた検定力分析

Statistical Power Analysis with Microsoft® Office Excel®

竹 安　大
（Hajime Takeyasu）

Ⅰ　はじめに

言語学の研究における数量的な観点を取り入れた研究が広がりを見せている。それに伴って、従来は実験音声学や心理言語学など一部の分野でのみ使われることが多かった統計学的分析が、音韻論、統語論、意味論など様々な分野で使用されるようになってきた。統計学的手法の使用に関する言語学者の立場は大きく 2 つに分けられる。すなわち、言語学では統計を使う必要などないとする立場と、積極的に統計学的手法を用いようとする立場である。ところが、後者には数学や統計学的手法になじみがないために統計的な分析をしたくてもどうしていいかわからないという者が少なくなく、言語学の分野における基礎的な統計学的知識の普及は急務の課題である。

このような流れの中、筆者は統計学の普及のための教育活動と勉強用に使える無償の分析ソフトの開発を続けてきた [1]。本稿では、言語学者にとってほぼなじみがないと思われ、かつ今後より需要が増すであろう検定力分析（検出力分析；Power Analysis）に関して、言語学の具体例（2 グループの平均値の比較）を用いながら、Cohen (1988)と Mathews (2010)で挙げられている近似法の正確さを統計ソフト R（R Development Core Team 2010、以下、単に R とする）による計算結果と比較し、近似法が非常によく正確であることを示したうえで、近似法による検定力分析がエクセル（Microsoft® Office Excel®）で簡単に実行できることを紹介する。

II　検定力分析

検定力分析は非常に重要であるにもかかわらず、分析方法に関するテキストはそれほど多くなく、また、手軽に検定力分析ができる統計分析用ソフトも数が限られている。本稿の目的は、特に言語学者向けに検定力分析の方法を紹介することである。この説では、まず、検定力分析とはどういうものなのかを言語学の例も交えながら簡単に紹介する。

II－1　検定力とは

検定力とは、帰無仮説が正しくないときにそれを正しく棄却できる確率のことを指し、検定力分析とは、その検定力を求める分析のことである [2]。帰無仮説が正しくないときに誤って帰無仮説を採択してしまう誤り（Type II error）をβとすると、検定力は$1-\beta$と表せる。帰無仮説が正しいときに誤って帰無仮説を棄却してしまう誤り（Type I error、一般にαと置かれる）に対して、β（または$1-\beta$）は補完的な関係にあると見なすことができる。なお、一般的にはαは0.05、βは0.20（すなわち、検定力は1－0.20＝0.80）を基準として設定されることが多いが、Cohen (1988)などでも述べられているように、これは慣習的なものであって絶対的なものではない [3]。

表１　帰無仮説の採択・棄却と Type I error、Type II error の関係 [4]

真実（母集団）＼検定結果	帰無仮説を棄却 （有意差あり； 差があると判断）	帰無仮説を採択 （有意差なし； 差がないと判断）
帰無仮説が正しくない （本当は差がある）	○ （検定力）	× （β：Type II error 率）
帰無仮説が正しい （本当は差がない）	× （α：Type I error 率）	○

検定力が特に大事な役割を果たすのは、実験計画における標本数の決定の場面や、実験実施後の有意差検定において帰無仮説が棄却されなかった場合（すなわち、有意確率 p が定められた基準（0.05 など）よりも大きかった場合）において特に重要である。例えば、検定力と標本数の間には、他の条件が同じであれば、標本数が大きくなるほど検定力が増すという関係が成り立つので、実験計画において検定力を高く設定すればするほど多くの標本を取らなければならなくなる。標本数の決定は実験計画において一般に重要な項目であるので、これに直結する検定力を理解することは非常に重要である。また、実験後の有意差検定において帰無仮説が棄却されなかった場合、多くの研究者が対立仮説は誤りであるという判断を下すが、これは検定力が非常に高い場合にのみ可能な解釈である。検定力は対立仮説が正しいときに、正しく帰無仮説を棄却できる確率であるから、検定力が低い場合には偶然（または標本数決定を含む実験計画上の問題で必然的に）帰無仮説を棄却できなかった可能性が大きいと見なされるためである。検定力が有意差検定を補完するものであるというのはこのような性質による。

II－2　言語学の具体例による検定力分析の例

英語の/s/と/ʃ/の子音持続時間が異なるかどうかを調べるため、英語話者 1 名に無意味語 /pabas/と/pabaʃ/を 10 回ずつ発音してもらったという状況で、/s/の平均持続時間が 88 msec. (SD = 7)、/ʃ/の平均持続時間が 87 msec. (SD = 10)だったと仮定する [5]。このデータについて、/s/と/ʃ/の平均持続時間は等しいという帰無仮説を検証するため、対応のない 2 グループに対する t 検定を行ったところ $t_{18} = 0.259$, $p = 0.799$ という結果が得られたとする。このとき、p の値は「帰無仮説（s の平均持続時間と ʃ の平均持続時間の差が 0 である）が正しいとき、今回の実験と全く同じことを繰り返し行ったときに今回のデータで得られたのと同程度の平均の差（88 msec.－87 msec.＝1 msec.）またはそれ以上の平均の差が偶然得られる確率は 79.9%である」ということを示しており、事

前に設定された Type I error の値（一般には 0.05）と比較して得られた p の値が高いことから、帰無仮説を棄却できない（有意差がない）という結論になる。ここまでの議論は、すべて Type I error 率、すなわち本当は差がないのに誤って差があると結論付けてしまう危険性に関連したものである。

さて、今回のように検定の結果有意差がなかった場合、多くの人は英語の/s/と/ʃ/の平均持続時間には差がないと述べてしまうが、この段階ではそのような結論を出すことはできない。繰り返しになるが、「帰無仮説を棄却できない」という場合、本当に差がない[6]という可能性と、本当は差があるのにそれを検知できず、誤って差がないと結論付けてしまった（Type II error）という可能性の 2 つがあるためである。そこで、III 節で紹介する Mathews (2010)の方法で検定力分析を行ってみたところ、検定力は 0.041 であった[7]。検定力が 0.041 であるということは、β、すなわち Type II error 率は 1－0.041＝0.959 であるから、実際には差があるのに誤って差がないと判断してしまうことにつながる可能性が極めて高いことになる（本当に差があるのかどうかは、標本数を増やすなどして検定力を高めたうえで議論する必要がある[8]）。よって、今回のデータからは/s/と/ʃ/の平均持続時間に有意差がないと断言することはできない。

III　2 グループの平均の差の分析における検定力の計算方法

2 グループの平均の差の分析における検定力は、以下に挙げる計算により求めた標準正規分布の累積分布関数の値（$z_{1-\beta}$）または Student の t 分布の値（t_{β}）を確率に変換することで得ることができる。ここでは、Cohen (1988)に挙げられている方法と、Mathews (2010)に挙げられている近似法を紹介する[9]。

III－1　Cohen (1988)の近似法

(1)の式によって得られた値を、標準正規分布の累積分布関数のパーセンタイ

ル値に変換することで検定力を求めることができる。なお、検定力は $z_{1-\beta}$ の値が大きくなるほど上がっていくから、(1)の式から、他の条件が同じであれば、d が大きくなるほど、また、（これはぱっと見ただけでは分かりにくいかもしれないが）標本数 n が大きくなるほど、検定力が増すことが分かる。

(1)

$$z_{1-\beta} = \frac{d(n-1)\sqrt{2n}}{2(n-1) + 1.21(z_{1-\alpha} - 1.06)} - z_{1-\alpha}$$

ただし、α は片側の値（よって、両側確率を求めたい場合は α を 2 で割る必要がある）、d は効果量[10]、n は各グループの標本数[11]、$z_{1-\alpha}$ は標準正規分布の累積分布関数の 1－α パーセンタイルの値とする。

III－2　Mathews (2010)の近似法（2 グループの標本数が等しい場合）

(2)および(3)の式によって求めた値から、t 分布の外側の部分の割合に変換[12]したものが β となるので、1－β によって検定力を求めることができる。t_β の値が大きくなるほど、β の値は小さくなるので、検定力は t_β の値が大きくなるほど上がっていくという関係にある。なお、(2)と(3)の式中の（　）の内の部分は Cohen (1988)の効果量 d の計算式と等しいので、Cohen (1988)の式のときと同様、効果量 d が大きくなるほど、また、標本数 n が大きくなるほど、検定力が増すことになる[13]。

(2) 対応のない 2 グループの場合

$$t_\beta = \sqrt{\frac{n}{2}}\left(\frac{\Delta\mu}{\sigma_\epsilon}\right) - t_{\alpha/2}$$

ただし、α は片側の値、n は各グループの標本数、$\Delta\mu$ は 2 グループの平均の差の絶対値、σ_ϵ は 2 グループの標準偏差、$t_{\alpha/2}$ は自由度 $2(n-1)$のときのスチューデントの t 分布（両側）の値とする。

(3) 対応のある 2 グループの場合 [14]

$$t_\beta = \sqrt{n}\left(\frac{|\overline{\Delta x}|}{\sigma_{\overline{\Delta x}}}\right) - t_{\alpha/2}$$

ただし、n は標本数（ペアの数）、$\overline{\Delta x}$は対になったデータの差の平均、$\sigma_{\overline{\Delta x}}$は対になったデータの差の標準偏差、$t_{\alpha/2}$は自由度が $n-1$ のときのスチューデントの t 分布（両側）の値とする。

以上に挙げた式からわかるように、検定力は効果量 d、標本数 n、α（事前に設定された Type I error 率）に連動するものである。このうち、α は自由に設定できるものではあるが、言語学の分野においてはほぼ固定的に 0.05 が用いられているため、実質的には効果量 d と標本数 n が検定力を決める要因となっている。

IV　近似法の結果の評価

Cohen (1988)の方法も Mathews (2010)の方法もいずれも近似法であり、標本数が大きいほど正確な値が得られることになる。いずれも方法も、正確な値と非常に近い値が得られると言われるが、近似法の当てはまりのよさは条件によっても異なるため、様々な条件において具体的にどのくらい近い値が得られるのかを確認してみることは有益である [15]。

以下では、2 グループの平均の比較における検定力を Cohen (1988)と Mathews (2010)の近似法を使って求め、その値を同じ条件で統計分析ソフト R（正確な値が得られる）で分析した場合と比較した結果を報告する [16]。なお、α、効果量 d、標本数 n については、言語学での使用という状況を想定し、それぞれ α（両側）を 0.05、効果量 d を 0.05～3.00、標本数 n を 10, 30, 50, 100, 500, 1000 とおいた。結果は、対応のない 2 グループの場合（図 1 から図 6）と対応のある 2 グループの場合（図 7 から図 12）に分けて示した。

IV－1　対応のない 2 グループの平均の比較の場合

図 1～図 6 は、x 軸に効果量 *d* を、y 軸に検定力を取り、Cohen (1988)の方法による計算結果、Mathews (2010)の方法による計算結果、R による計算結果をプロットして作成した。各図は標本数が異なっており、図 1 から図 6 まで順に $n = 10, 30, 50, 100, 500, 1000$ という条件に対応している。

3 種類の方法による結果を比較すると、効果量 *d* と標本数 *n* がともに小さい場合には若干のずれが観察されるものの [17]、3 本の線はほぼ重なっており、いずれの方法でもほぼ同じ結果が得られると見なすことができる [18]。もちろん、正確な値からのずれをどの程度許容するかは一概には決められない部分があるが、少なくとも言語学やそれに関連する分野における使用という実用的観点からは、Cohen (1988)や Mathews (2010)でもそれぞれ述べられている通り、近似法の結果を用いても問題はないと言ってよいであろう。もちろん、ここで問題がないと述べたことは筆者の主観であり、筆者が属する研究分野（主に音声学・音韻論・実験心理学）では普通小数点第 3 位よりも下の値の違いまで細かく議論されることは少ないという経験に基づくものであるから、研究分野によってはこれは当てはまらないかもしれない。仮に自身の研究分野で非常に細かい値まで議論することが要求されるのであれば、状況によっては近似法の使用は避けるべきであることは付け加えておきたい。

計算式からも読み取れることであるが、図からは以下のようなことも読み取れる。まず、効果量 *d* の値が装荷するにつれて検定力が増加していくことがわかる。また、図 1～図 6 の違いは標本数 *n* の指定の違いにあるが、標本数が増えるほど分布が左寄りになっていく、すなわち、検定力が増加していくことも見て取れる。まとめると、以下のようになる。

(4) 他の条件が同じであれば、

a. 効果量 *d* が増加すると検定力が高くなる

b. 標本数 *n* が増加すると検定力が高くなる

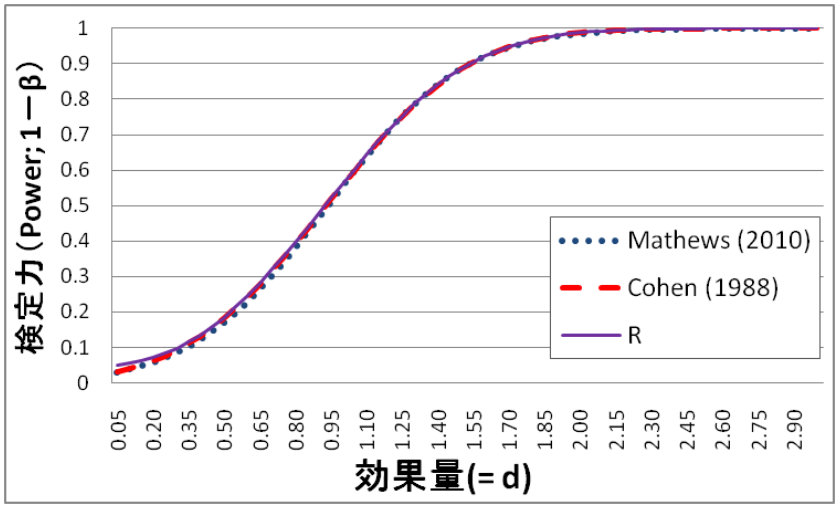

図1: 対応のない 2 グループの検定における $\alpha = 0.05$, $n = 10$ のときの効果量 d と検定力の関係

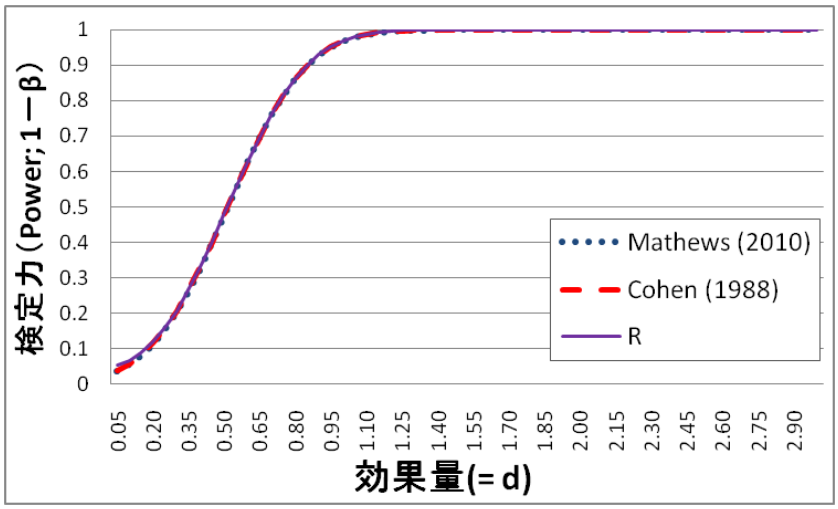

図2: 対応のない 2 グループの検定における $\alpha = 0.05$, $n = 30$ のときの効果量 d と検定力の関係

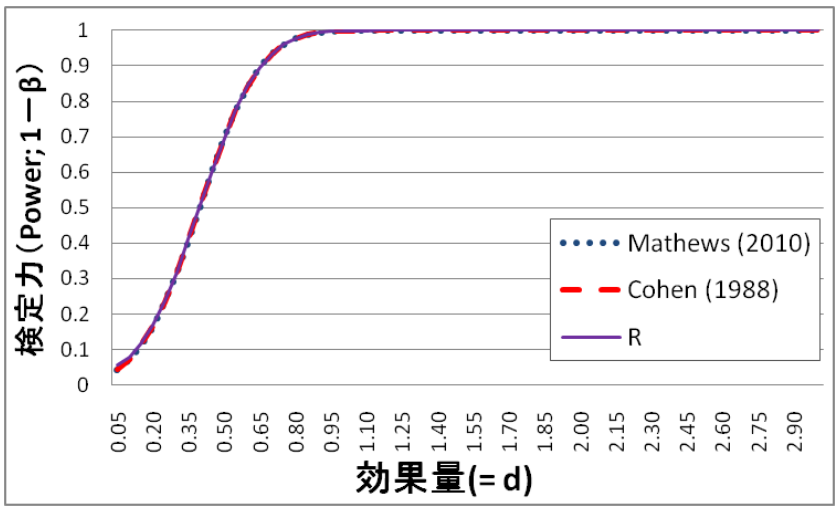

図3: 対応のない 2 グループの検定における $\alpha = 0.05$, $n = 50$ のときの効果量 d と検定力の関係

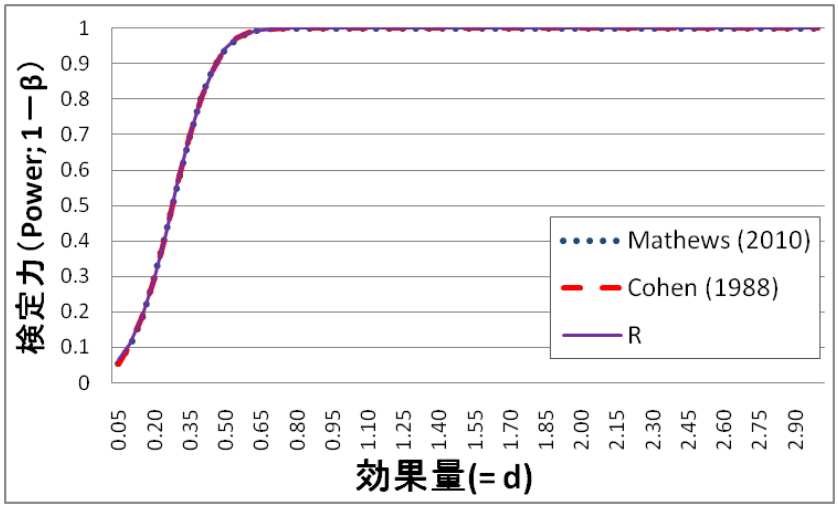

図4: 対応のない 2 グループの検定における $\alpha = 0.05$, $n = 100$ のときの効果量 d と検定力の関係

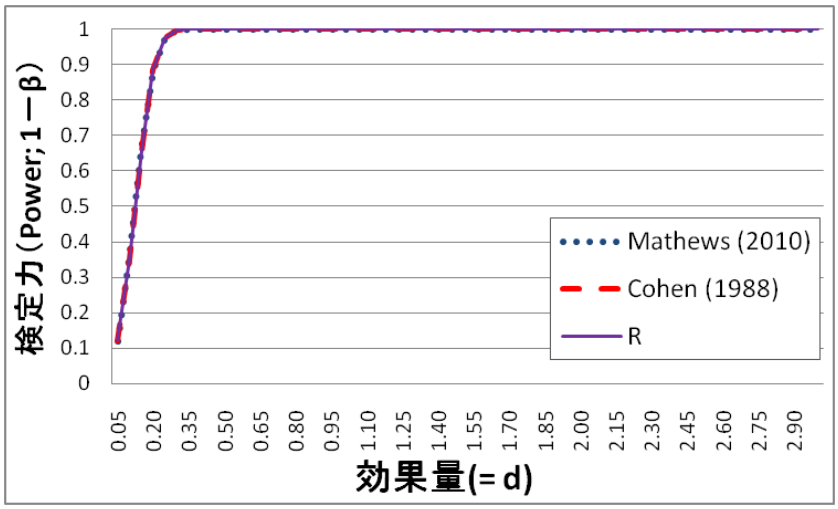

図5: 対応のない 2 グループの検定における $\alpha = 0.05$, $n = 500$ のときの効果量 d と検定力の関係

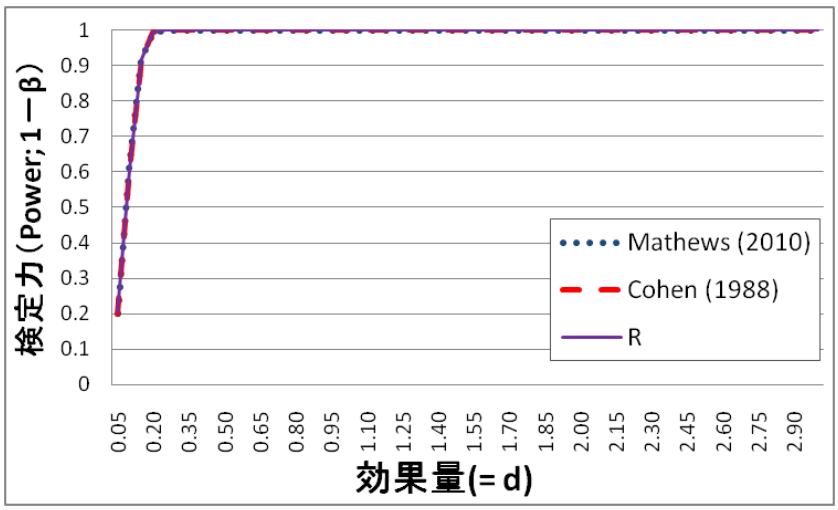

図6: 対応のない 2 グループの検定における $\alpha = 0.05$, $n = 1000$ のときの効果量 d と検定力の関係

IV－2　対応のある 2 グループの平均の比較の場合

図 7～図 12 は対応がある 2 グループに関する結果を示したものである。対応がある 2 グループの場合、標本数 n が少ない（今回であれば、n＝10）のときには計算方法の違いによって多少のずれが見られる。しかし、一般的な傾向は対応がない 2 グループの場合と同様で、効果量 d が増加するほど、また、標本数 n が増加するほど検定力も高くなるという関係がある。

また、これは本稿の議論とは直接関係がないことであるが、同じ規模の実験であれば一般に対応がある 2 グループの比較の方が対応がない 2 グループの比較に比べて効率が良いと言われており（Mathews 2010）、それをここに挙げた図から読み取ることができる。例えば、図 1 と図 7 を比較してみると図 7 の方が結果が左寄りに分布していることから、条件が同じであれば、対応がある 2 グループの方が対応がない 2 グループの場合よりも検定力が高くなっていることが分かる。図 2 と図 8、図 3 と図 9、図 4 と図 10、図 5 と図 11、図 6 と図 12 にもそれぞれ同様の関係がある。標本数 n が同じであれば対応がある 2 標本の方が高い検定力が得られるということは、同じ程度の検定力を得るために必要な標本数は対応がある 2 グループの方が少なくて済む（効率がいい）ということになるわけである。

なお、ここで言う「標本数 n が同じ」という点について一言付け加えておきたい。対応のない 2 グループにおける標本数 n とは、1 グループ当たりの標本数のことだから、n＝10 とは、必要な標本数は全部で 10×2＝20 標本であることになる。一方、対応のある 2 グループの場合、標本数 n はペア数に相当するから、何らかの基準に基づく対になった標本が 10 あればいいことになる。対応のある 2 グループでも、対になったデータを作るために観察数自体は 10×2＝20 が必要なわけだが、標本数が被験者数に相当するような状況では、20 名集める必要があるか 10 名集めるだけでよいかという違いになってくる。こうした特徴を理解しておけば実験実施の労力を抑えることも可能となるわけである。

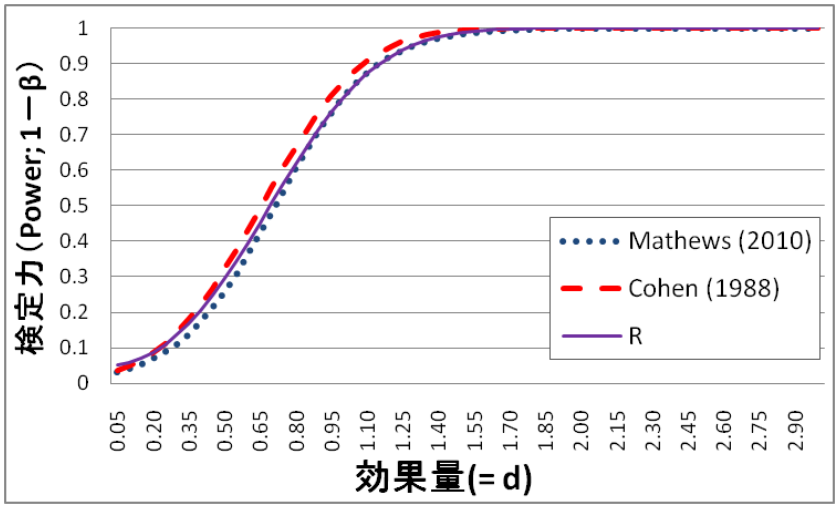

図7: 対応のある 2 グループの検定における $\alpha = 0.05$, $n = 10$ のときの効果量 d と検定力の関係

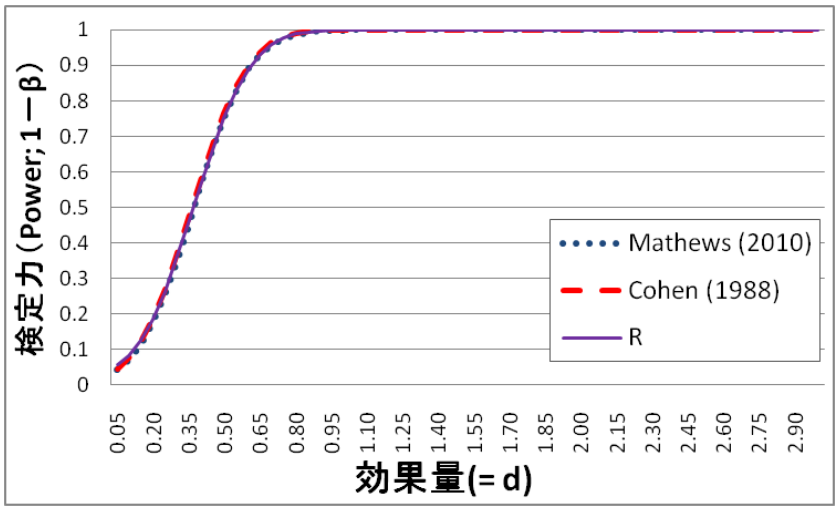

図8: 対応のある 2 グループの検定における $\alpha = 0.05$, $n = 30$ のときの効果量 d と検定力の関係

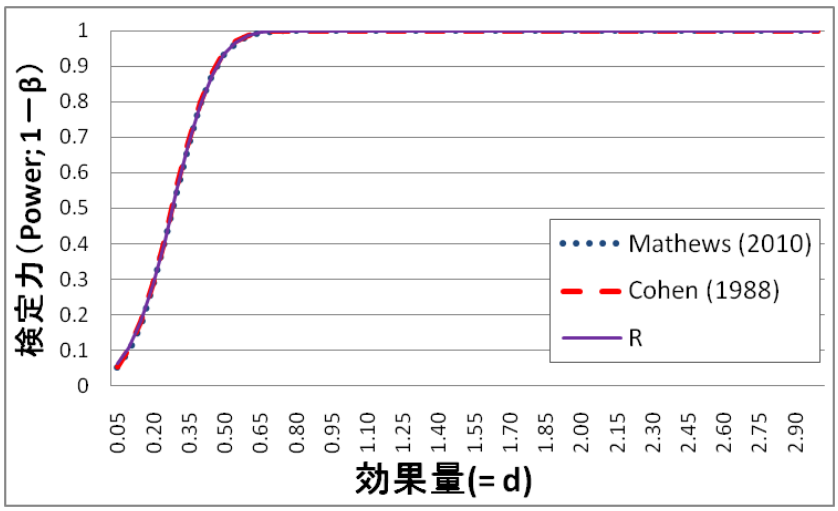

図9: 対応のある 2 グループの検定における $\alpha = 0.05$, $n = 50$ のときの効果量 d と検定力の関係

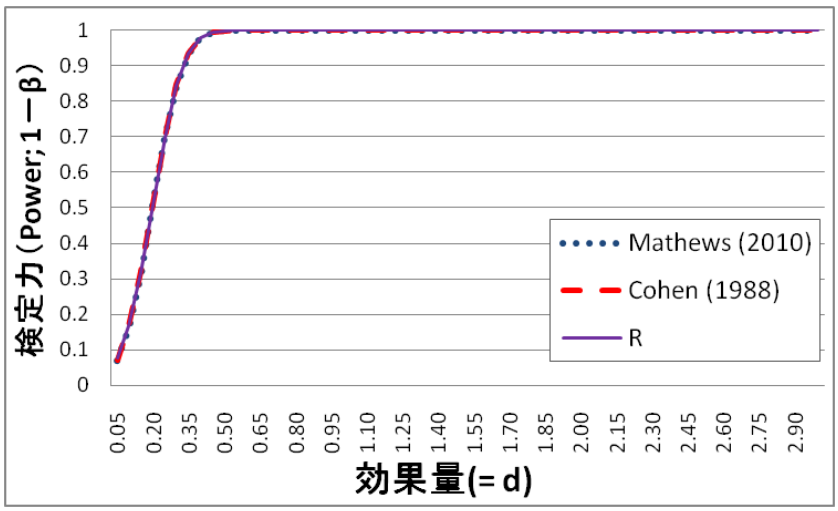

図10:　対応のある 2 グループの検定における $\alpha = 0.05$, $n = 100$ のときの効果量 d と検定力の関係

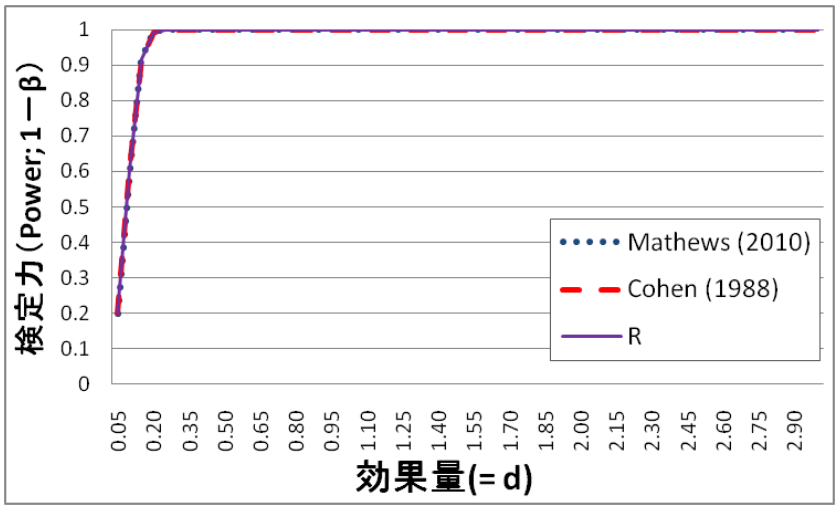

図11:　対応のある 2 グループの検定における $\alpha = 0.05$, $n = 500$ のときの効果量 d と検定力の関係

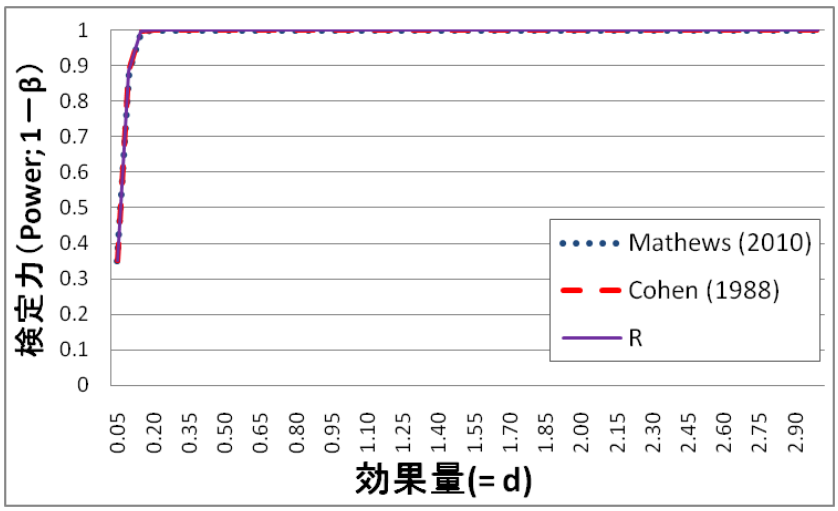

図12:　対応のある 2 グループの検定における $\alpha = 0.05$, $n = 1000$ のときの効果量 d と検定力の関係

Ⅴ　エクセルによる検定力分析

検定力分析は非常に重要であるにもかかわらず、分析方法に関するテキストはそれほど多くない。また、手軽に検定力分析ができる統計分析用ソフトも数が限られている。最後に、本稿で紹介した近似法をエクセルで簡単に行う方法を紹介する。

エクセルを使って足し算や掛け算などの単純な計算をする方法はここでは省略するが、それらがわかっていれば、以下のような統計学的関数を使って(1)～(3)の式を計算することができる。以下は使用すべき関数のリストである。いずれも、(　)内には各項目について指定したい数値を入れる必要がある。また、但し書きの部分はエクセルの関数を使う時に注意すべき点である。

(5) 数式中の項目の計算のために使う関数

- $z_{1-\alpha}$：NORMSINV(1－α)
  ただし、両側の値を求めたい場合、αには$\alpha \div 2$ の値を入れる必要がある。
- $t_{\alpha/2}$：TINV(α, 自由度)
  NORMSINV と異なり、両側の値を求めたい場合はαの値は 2 で割らずにそのまま入れるとテキスト（Mathews 2010）の通りの結果となる。

(6) 計算結果を確率に変換するのに使う関数

- $z_{1-\beta}$を確率に変換する場合、NORMSDIST($z_{1-\beta}$で得られた値)を用いる。この結果得られた値が 1－β、すなわち検定力となる。
- t_{β}を確率に変換する場合、TDIST(t_{β}で得られた値, 自由度, 1)を用いる。この結果得られた値がβになるので、1 からβを引いた値が検定力となる。

検定力分析の機能は様々な統計分析用ソフトに組み込まれており、有料のものも多いが、中には本稿の分析でも用いた R のように無料で使用できるものもある[19]。また、統計分析用ソフトでなくても、エクセルのような身近な汎用ソ

フトを使って以上で議論してきたCohen (1988)とMathews (2010)の近似法を実行することもできる（(1)～(3)の式を入力するだけで可能[20]）。検定力分析を行ってみたいが計算式を入力するのは面倒だという人向けに、必要項目（α、標本数、効果量）だけ記入すれば自動で(1)～(3)式の計算結果を表示するようにエクセルで組んだ検定用フォーマットが筆者のホームページ（「サイエンスカフェ：言語学のための統計」http://sites.google.com/site/hponsei/sci-cafe）に置いてあるので、実際に分析してみたいという方は使ってみていただければ幸いである[21]。

VI　まとめ

本稿では、今後様々な分野で使用される機会が増すであろう検定力分析に関して、言語学の具体例（2 グループの比較）を用いながら、Cohen (1988)とMathews (2010)で挙げられている近似法の正確さを統計ソフト R による計算結果と比較し、近似法が非常によく正確であることを示した。また、近似法による検定力分析がエクセルで簡単に実行できることを紹介した。

注

1. 詳細は「サイエンスカフェ：言語学で使える統計」ホームページ（http://sites.google.com/site/hponsei/sci-cafe）を参照のこと。
2. 本稿の議論は有意差検定に関する知識を前提としている。有意差検定に関する知識は上述の「サイエンスカフェ：言語学で使える統計」ホームページや、一般的な統計学のテキストから得ることができる。また、本稿では主に有意差検定における検定力分析とその近似法計算を議論するが、有意差検定は信頼区間とも密接に関係しているので、本稿での議論は信頼区間の議論にも当てはまるものである（詳しくは Mathews 2010 などの統計学のテキストを参照のこと）。
3. 検定力がどの程度高ければいいのかは状況によっても異なるが、有意差があると主張したい場合、実験計画時点では既述の通り検定力が 0.8 程度になるように標本数を決めるのが一般的なようである。判断が難しいのは有意差がないと主張したい場合で、Type I error 率（α）を 0.05 に設定するのと並行して、Type II error 率（β）も 0.05 にするた

めに検定力を 0.95 に設定するというのも一つの手ではあるが、差がないことを主張する場合には想定する効果量も小さくなるはずなので、検定力の基準をあまりに高くしてしまうと膨大な数の標本が必要とされ、実験実施に多大な労力を要することになってしまう。このような場合には、実験実施のコストと結論が出せない場合に生じるコストを比較したうえで折り合いをつけるしかない。

4. より詳しい説明は、筆者が 2010 年 3 月に行った第 15 回サイエンスカフェの配布資料「より適切に有意差検定を用いるために」（同ホームページよりダウンロード可能）も参照のこと。
5. このデータは、筆者の博士論文（竹安 2009）で実際に扱ったデータの一部である。
6. ここでの「差がない」は、無視できるほどに小さな差であるという場合も含むものとする。
7. なお、同じデータに対して Cohen (1988)の方法を用いた場合の検定力は 0.043、統計分析ソフト R を用いた場合は 0.057 であった。後ほど述べるように、標本数が少なく効果量が小さい場合、Mathews (2010)や Cohen (1988)の近似法は正確な値（ここでは 0.057）から多少ずれる。
8. この例は、実験計画における検定力分析（必要標本数の決定を含む）を怠った結果、データを取ってみたものの結論が出せないという典型例であり、実験計画の重要性を示しているものである。ただし、今回のデータは結論を出すには遠いものの、予備的実験と見なして今後の実験に役立てていくことは可能である。すなわち、ここで得られた効果量を母集団推定値と考えて、必要な標本数推定に役立てるのである。ちなみに、今回の/s/, /ʃ/のデータにおける効果量の値を用いて、検定力を 0.8 にするために必要な標本数を計算した結果、/s/, /ʃ/ それぞれについて 1170 ずつ、合計 2340 トークンのデータが必要になることがわかった（さらに検定力を上げようと思えばそれ以上の標本数が必要となる）。差がないことを強く主張することが非常に大変なことであることがわかる。
   なお、次節に挙げた(1)～(3)式を移項して n に関する式に書き直せば必要標本数を求めることができるが、必要標本数の決定に関する細かい説明や具体例による説明は別の機会に譲ることとしたい。
9. Cohen (1988)および Mathews (2010)はいずれも統計学の教科書なので、これらの方法はそれぞれ Cohen (1988)と Mathews (2010)が独自に考案したものだというわけではないはずであるが、ここでは原典を明らかにするのが目的ではないので以上 2 つの文献を挙げている。
10. 効果量は Cohen (1988)の定義によるもので、対応のない 2 グループの差の比較の場合、平均の差÷σ’（2 グループの標準偏差）、対応のある 2 グループの差の比較の場合、ペアごとの差の平均÷ペアごとの差の標準偏差となる。なお、A, B をそれぞれグループを表すものだとすると、σ’は以下の式により求めることができる。

$$\sigma' = \sqrt{\frac{\sigma_A^2 + \sigma_B^2}{2}}$$

11. 2 グループに対応がない場合、グループ①の標本数が 10、グループ②の標本数が 10 だとすると、ここで n に入れるべき値は 20 ではなく 10 となる。グループ①と②の標本数が異なる場合には、Cohen (1988)の修正式が挙げられているのでそれを用いることができる。2 グループで対応がある場合には、n はペア数となる。
12. t 検定で t の値を p に変換するときにするのと同じ方法。
13. Cohen (1988)は効果量の使用を強く勧めているが、Mathews (2010)は Cohen の定義による効果量は使用すべきでないと述べている。これはどちらが正しいというような種類のものではなく、単に両者の意図の違いにあるものと考えられる。Cohen は有意差検定偏重の時代背景の中で、また、メタ分析的な観点を視野に入れた議論の中で効果量の使用を強く勧めているのに対し、Mathews (2010)は特に実験計画における標本数の計算を考慮したときに効果量を平均と標準偏差を分けて考えておくほうが様々な点で有益だというニュアンスで Cohen の効果量の使用を否定していると思われるためである。Cohen の効果量の使用の是非はともかくとして、Mathews (2010)による計算式中の平均÷標準偏差は Cohen の効果量 d と等しいので、本稿では以降も効果量 d を説明に用いる。
14. 対応のある 2 グループに関する式は、Mathews (2010: 31)の(2.21)式を対応がある 2 グループと比較がしやすい形に書き換えて示してある。
15. 近似法の当てはまりの良さに関しては、様々な文献で議論されているものと思われるが、多くは言語学者が目にする機会がないであろうと思われる分野の研究論文として出されており、仮に目にしたとしても読み解くことが困難であることが推測されるので、ここで近似法の結果の比較をすることには意味があると判断した。
16. R の分析では Stéphane Champely による pwr.t.test のパッケージを用いた。
17. 近似法の場合、グループ数が大きくなるほど正確な値への近似の度合いが改善されるので、n の値が増えるほど 3 本の線がより重なっていくのは自然なことであるが、標本数 n が 10 という少なめの値であってもほとんどずれが生じていないことは、言語学者にとっては喜ばしい結果であると言えるかもしれない。というのも、言語学の分野では比較的少ない標本に基づいて議論が展開されることが多く（これはあくまで筆者の個人的な印象ではあるが、例えば音声学の実験であれば「標本数が 10 あればとりあえず安心」といった暗黙の了解のようなものがあるように思われる）、近似の度合いの問題が生じやすい環境であると思われるためである。
18. ここで注意しなければならないのは、これはあくまで計算結果の正確性の議論であるということである。図からわかるのは標本数が少なくても近似法によって正確な値に近い計算結果が得られるということであって、個々の研究者が行う少ない標本数に基づいた

実験やその結果に基づく主張が妥当であるかどうかとは全く別の事柄である。

19. 使いやすさはソフトによって大きく異なっており、R のように専用のコマンドを覚えないと動かせないものから、ワークシート形式になっていて使いやすいものまで様々であるが、使いやすいものはたいてい有料（非常に高額）である。なお、R のコマンド入力は慣れている人からすると大して難しくはない事柄ではあるかも知れないが、統計学に触れたばかりの初心者の感覚としては非常に困難な作業である。実際、筆者が出会ってきた言語学者の多くは新しい入力方法を覚える必要があるとなった時点で、どうしても使わなければならないという事情がない限りソフトの使用を放棄してしまう場合がほとんどであった。中には、自分が知らないソフトを新たにインストールすることさえ面倒だと感じる人もいるほどである。言語学者に気軽に使ってもらうためには、無料でかつ操作が非常に簡単という条件が必須である。
20. ただし、エクセルの統計的関数は Cohen (1988)や Mathews (2010)の定義と少し異なる部分があるため、その点の修正は必要である。
21. 2011 年 2 月の時点で、「t 検定・相関分析」のシートに、2 グループの平均の比較に関して Cohen (1988)および Mathews (2010)の方法による計算結果を表示する機能がついている。また、検定力分析と、実験計画における必要標本数計算に特化したエクセルファイルも近日アップする予定なので、興味がある方はホームページをチェックしてもらえたら幸いである。

参考文献

Cohen, Jacob (1988). *Statistical Power Analysis for the Behavioral Sciences* (2nd ed.). New York: Psychology Press.

Mathews, Paul (2010). *Sample Size Calculations: Practical Methods for Engineers and Scientists*. Ohio: Mathews Malnar and Bailey, Inc.

R Development Core Team (2010). R: A language and environment for statistical computing. R Foundation for Statistical Computing, Vienna, Austria. ISBN 3-900051-07-0, URL http://www.R-project.org/.

竹安大（2009）「音韻的有標性とその音声学的基盤」神戸大学大学院文化学研究科博士論文.（神戸大学学術成果リポジトリ Kernel : http://www.lib.kobe-u.ac.jp/repository/thesis/d1/D1004803.pdf）

竹安大・秋田喜美（2011）「サイエンスカフェ：言語学で使える統計」検定用フォーマット (Version: 20112024), Microsoft Excel ファイル. (URL: http://sites.google.com/site/hponsei/sci-cafe)